무선 엣지 네트워크에서 견고하고 효율적인 연합학습을 위한 불확실성 기반 단말 스케줄링 기법

박기태, 홍준표
동국대학교, 홍익대학교
7045rlxo@dgu.ac.kr, jp_hong@hongik.ac.kr

Uncertainty-Aware Scheduling for Robust and Efficient Federated Learning over Wireless Edge Networks

Gitae Park, Junpyo Hong
Dongguk Univ., Hongik Univ.

요 약

본 논문에서는 단말들이 보유한 데이터가 희소하고 이질적인 무선 엣지 네트워크 환경에서, 학습 편향과 제한된 통신 자원으로 인해 학습이 원활이 진행되지 않는 기존 연합학습의 한계를 극복하기 위해 베이지안 접근을 활용한 업데이트 중요도-채널 기반 스케줄링을 제안한다. 제안방안에서는 각 단말이 로컬 데이터셋을 활용해 변분 추론(variational inference) 기반 베이지안(Bayesian) 학습을 수행해 모델 파라메터에 대한 불확실도를 정량화하고, 이를 기반으로 업데이트 중요도를 도출해 지역 업데이트를 전송할 단말 스케줄링에 무선 채널 상태와 함께 고려한다. 이를 통해 유용한 지역 업데이트를 선별적으로 학습에 반영함으로써, 전체 학습시간과 통신 비용을 최소화하면서도 열악한 환경에서 모델의 성능을 효과적으로 향상시킬 수 있다. 시뮬레이션 결과, 제안한 중요도 기반 스케줄링 기법이 기존 방식 대비 통신 효율과 학습 수렴 속도 측면에서 우수한 성능을 보임을 확인하였다.

Ⅰ. 서 론

최근 연합학습(federated learning, FL)은 프라이버시 보호와 IoT기기의 확산으로 주목받고 있다. 특히, 5G/6G 엣지 환경에서 무선 지연, 자원 제약이 연합학습 성능의 핵심 변수가 된다.

기존 연구들은 모델 파라미터를 고정된 상수로 간주하고 관측 데이터를 가장 잘 나타내는 단일 추정지를 도출하는 빈도주의적 접근(frequentist approach)을 바탕으로 FedAvg[1]을 확장하여 전송 지연, 에너지 효율, 이기종(heterogenous) 자원 제약을 최적화하는 클라이언트 스케줄링 기법을 제안해 왔다. 대표적인 예로 FedCS[2]는 계산, 통신 자원을 반영해 클라이언트를 스케줄링한다. 하지만 빈도주의 FL은 클라이언트들이 보유한 데이터가 희소하고, 분포가 상이할 경우, 서로의 지역 업데이트가 크게 달라 학습이 잘 진행되지 않는 문제를 갖는다.

이를 극복하기 위해, 최근 베이지안(Bayesian) 방법론이 연합학습에 도입되었다. 이는 모델의 사후분포를 도출해 다양한 모델들에 대한 가능성을 가습과정에 고려함으로써 기존 빈도주의 연합학습이 갖는 문제들을 완화할 수 있었다.

본 논문은 이러한 베이지안 관점에서, 각 단말의 업데이트 중요도와 무선 채널 이득을 융합한 중요도 기반 단말 스케줄링 기법을 제안한다. 이를 통해 전체 통신 지연을 최소화하고 학습 수렴 속도를 최대화하는 것을 목표로 한다.

Ⅱ. 시스템 모델

본 장에서는 무선 베이지안 연합학습이 수행되는 단말-서버 구조, 채널-전송 모델, 최적화 문제를 정의한다. 네트워크는 서버(Base station, BS) 1대와 $K \in \mathcal{K} = \{1, ..., K\}$ 대의 단말로 이루어져 있다. 학습은 총 T번의 라운드로 이루어진다. 베이지안 연합학습의 목표는 글로벌 사후분포 $p(w|D)$를 추정하는 것이다. 여기서 $D = (D_1, ..., D_K)$는 데이터셋, w는 모델 파라미터를 나타낸다. 그러나 계산 복잡도로 인해 $p(w|D)$를 직접 구하는 것은 현실적으로 불가능하기 때문에 이를 근사할 수 있는 간단한 형태의 변분 사후분포 $q(w|\theta) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 학습한다. 여기서 최적 파라미터 $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$는

$$\min_{\theta} D_{KL}[q(w|\theta) \| p(w|D)], \quad (1)$$

를 해결함으로써 얻는다. 베이즈 정리를 적용하면 식(1)은

$$\min_{\theta} D_{KL}[q(w|\theta) \| p(w)] + E_{q(w|\theta)}[-\log p(D|w)], \quad (2)$$

와 동치이며, 두 번째 항은 스케줄링된 단말 집합 S_t의 단말별 데이터로 분해된다.

$$E_{q(w|\theta)}[-\log p(D|w)] = \sum_{k \in S_t} E_{q(w|\theta)}[-\log p(D_k|w)], \quad (3)$$

따라서 식(2)는 다음과 같이 쓸 수 있다.

$$\min_{\theta} D_{KL}[q(w|\theta) \| p(w)] + \sum_{k \in S_t} E_{q(w|\theta)}[-\log p(D_k|w)], \quad (4)$$

단말 k의 데이터셋 크기를 $|D_k|$라 할 때

$$\pi_k^{(t)} = |D_k| / \sum_{j \in S_t} |D_j| \quad \sum_{k=S_t} \pi_k^{(t)} = 1,$$

로 정의하면, 글로벌 손실 함수를 가중합 형태로 나타낼 수 있다.

$$L_{glob}(\theta) = \sum_{k=S_t} \pi_k^{(t)} \{E_{q(w|\theta)}[-\log p(D_k|w)] + D_{KL}[q(w|\theta) \| p(w)]\}, \quad (5)$$

제안기법의 라운드t 시작 시, 서버는 글로벌 사후분포 $q(w|\theta_t)$를 모든 단말에게 브로드캐스트한다. 단말 k는 로컬 손실 함수

$$L_k(\theta) = E_{q(w|\theta)}[-\log p(D_k|w)] + D_{KL}[q(w|\theta) \| q(w|\theta_t)], \quad (6)$$

를 최소화는 $\theta_{t,k} = (\boldsymbol{\mu}_{t,k}, \boldsymbol{\Sigma}_{t,k})$를 도출한다. 서버는 업로드된 지역 분포들의 곱연산을 통해 지역학습 결과를 집계하며, 이에 따른 글로벌 분포는 다음과 같이 갱신된다.

$$q(w|\theta_{t+1}) \propto \prod_{k=S_t} q(w|\theta_{t,k})^{\pi_k^{(t)}}, \quad (7)$$

식(7)은 식(5)를 정확히 최소화하는 최적 해임이 증명되어 있다[3].

로컬 업데이트는 TDMA 기반으로 전송되며, 라운드 t에서 단말 k의 수신 신호는 다음과 같다.

$$y_{t,k}[n] = h_{t,k}x_{t,k}[n] + z_{t,k}[n], \tag{8}$$

이때, $z_{t,k[n]} \sim \mathcal{N}(0,\sigma_z^2)$이다. 채널 이득 $h_{t,k}$는 quasi-static Rayleigh fading을 가정해 전송 중에는 상수이고, 라운드 종료 후 $\mathcal{CN}\left(0, d_{t,k}^{-\alpha}\right)$에 따라 갱신된다. 여기서 $d_{t,k}$는 라운드 t에서 단말 k와 서버간의 거리를 나타내며 α는 path loss exponent이다. 채널 상태 정보(channel state information, CSI)는 서버에서 알고 있다고 가정한다. 대역폭 W에 대해서 채널 모델의 용량은 다음과 같다.

$$C_{t,k} = W\log\left(1+\frac{|h_{t,k}|^2 P}{\sigma_z^2}\right), \tag{9}$$

이때, $P = E[|x_{t,k}[n]|^2]$이다. 모델이 d개의 파라미터를 갖고 각 파라미터는 b bits으로 양자화되었을 때, 베이지안 학습은 각 파라미터의 평균과 분산을 학습하므로 로컬 업데이트의 크기는 $B = 2db$ bits가 된다. 따라서 로컬 업데이트 전송에 걸리는 시간은 다음과 같다.

$$\tau_t = \sum_{k \in S_t} \frac{B}{C_{t,k}}, \tag{10}$$

모델의 수렴까지 T번의 라운드가 필요하다면, 전체 소모 시간은 다음과 같다.

$$\tau_{total} = \sum_{t=1}^{T} \tau_t, \tag{11}$$

라운드마다 일부 단말만 선택하면 업로드 지연은 줄어들지만, 의미가 없는 지역 업데이트를 가진 단말이나 잘못된 방향으로 학습을 유도할 수 있는 단말이 스케줄링 될 경우, 오히려 수렴에 필요한 라운드 수가 늘어날 수 있다. 따라서 채널 조건과 변분 사후분포 기반 파라미터 신뢰도를 통합한 중요도 기반 스케줄링으로 총 학습시간 τ_{total}의 최소화를 목표로 한다.

Ⅲ. 중요도 기반 스케줄링

지역 학습 후, 단말 k는 다음에 따라 라운드 t에서의 업데이트 중요도 $\gamma_{t,k}$를 계산한다.

$$\gamma_{t,k} = \sqrt{\boldsymbol{\mu}_{t,k}^T \boldsymbol{\Sigma}_{t,k}^{-1} \boldsymbol{\mu}_{t,k}} = \sum_{i=1}^{d} \frac{|\mu_{t,k,i}|}{\sigma_{t,k,i}}, \tag{12}$$

이때, $\mu_{t,k,i}$는 $\boldsymbol{\mu}_{t,k}$의 i번 째 대각원소, $\sigma_{t,k,i}$는 $\boldsymbol{\Sigma}_{t,k}$의 i번 째 대각원소의 제곱근이다. $\gamma_{t,k}$는 파라미터 불확실성이 낮을수록 크게 정의하였다. 구체적으로, 평균의 절댓값이 클수록 추정이 뚜렷하고, 표준편차가 작을수록 추정 오차가 작으므로 중요도가 증가한다. 서버는 채널 정보를 바탕으로 각 단말별 채널 품질 지표 $g_{t,k}$를 구한다.

$$g_{t,k} = |h_{t,k}|^2 / d_{t,k}^{-\alpha}, \tag{13}$$

채널 품질 지표 $g_{t,k}$는 $d_{t,k}^{-\alpha}$로 정규화하여 소규모 페이딩만을 반영하므로, 스케줄링 단계에서 단말 간 거리 편차를 제거하고 순간적인 페이딩 우위만을 이용해 공정한 스케줄링 효과를 만들어 낸다.

각 단말은 로컬 학습이 끝난 후 $\gamma_{t,k}$를 계산하여 서버에 업로드하고, 서버는 $g_{t,k}$를 계산한다. 그 후 서버에서 $\gamma_{t,k}$가 높은 상위 x%의 단말 중 $g_{t,k}$가 높은 순으로 $\hat{K}$개 단말을 스케줄링 한다.

Ⅵ. 실험

기본 파라미터는 $K = 40$, $\hat{K} = 7$, $x = 50$, $W = 1\text{MHz}$, $P = 20\text{dBm}$, $\sigma_z^2 = -70\text{dBm}$, $\alpha = 4$로 설정하였다. 단말은 BS 반경 200m 내에 랜덤 배치하였으며, 로컬 데이터 크기는 $|D_k| \sim \text{Pois}(10)$로 생성하였다. 최대 라운드는 $T = 1000$으로 제한하였다. 모델은 3-layer CNN, 데이터셋은 MNIST, 베이지안 네트워크는 Blitz[4]로 구현했으며, 동일 구조의 빈도주의 모델을 비교군으로 사용하였다. 비교방안의 스케줄링에 gradient 기반 중요도 및 채널 기반 스케줄링을 적용하였다[5].

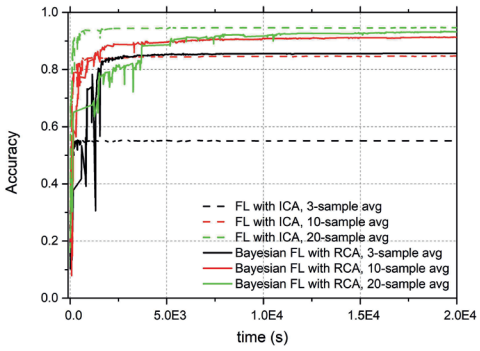

그림 1. 평균 샘플 수에 따른 수렴 그래프

Fig. 1. Convergence graph with respect to average number of samples

그림 1은 전체 단말의 평균 샘플 수에 따른 수렴 그래프이다. 기존 연합학습의 경우, 각 단말의 보유 데이터 샘플 수가 평균 3개일 때 정확도가 0.55 정도로 수렴하며 샘플 수가 10개는 되어야 정확도가 0.85 수준에 수렴하는 등 데이터 부족 시 비효율적인 단말 선택으로 수렴이 지연된다. 반면 제안방안은 사후분포의 신뢰도와 채널 상태를 함께 고려해 단말 당 평균 샘플 수가 3으로 적은 환경에서도 0.83이상의 높은 정확도를 보이며 기존방안보다 최대 30%높은 정확도를 보인다. 또한 샘플 수가 10개인 경우 같은 환경의 기존방안에 비해 더욱 높은 정확도에 수렴한다. 이에 더해 데이터가 20개로 충분한 상황에서도 기존방안과 비슷한 정확도에 수렴하는 것을 확인할 수 있다. 이를 통해 베이지안 중요도 지표가 제한된 무선 환경에서 효과적으로 작동함을 알 수 있다.

ACKNOWLEDGMENT

참 고 문 헌

[1] McMahan, Brenden, et al. “Communication-efficient learning of deep network from decentralized data,” *Artificial intelligence and statics*. PMLR, 2017.

[2] Nishio, Takayuki, and Ryo Yonetani. “Client selection for federated learning with heterogeneous resources in mobile edge,” *ICC 2019-2019 IEEE international conference on communication (ICC)*. IEEE, 2019.

[3] AI-Shedivat, Maruan, et al. “Federated learning via posterior averaging: A new perspective and practical algorithms,” *arXiv preprint arXiv:2010.05273* (2020).

[4] Esposito, Piero. “Blitz-bayesian layers in torch zoo (a bayesian deep learning library for torch).” *URL: https://github.com/piEsposito/blitz-bayesian-deep-learning* (2020).

[5] Ren, Jinkel, et al. “Scheduling for cellular federated edge learning with importance and channel awareness,” *IEEE Transaction on Wireless Communications* 19.11 (2020): 7690-7703.